Constitution d'un corpus de français tchaté

Falaise Achille

RÉCITAL 2005, Dourdan, 6-10 juin 2005

Constitution d'un corpus de français tchaté

Achille Falaise

GETA, CLIPS-IMAG – UJF - Université Grenoble I
385, rue de la Bibliothèque, B.P. 53, 38041 Grenoble Cedex 9
achille.falaise@imag.fr
Date prévue de la thèse : janvier 2008

Mots-clefs – Keywords

langue tchatée, ressources linguistiques, collecte de données

chat language, linguistic resources, resource acquisition

Résumé – Abstract

Nous présentons dans cet article un corpus de français tchaté, destiné à l'étude de la langue du tchat. Ce corpus, collecté et encodé automatiquement, est remarquable avant tout par son étendue, puisqu'il couvre un total de 4 millions de messages sur 105 canaux, hétérogènes sur les plans thématique et pragmatique. Son codage simple ne sera toutefois pas satisfaisant pour tous les usages. Il est disponible sur un site Internet, et consultable grâce à une interface web.

We present in this article a french chat corpus, intended for the study of chat language. This corpus, automatically collected and coded, is especially remarkable for its extent, since it covers a total of 4 million messages on 105 channels, heterogeneous from a thematic and pragmatic point of view. Its simple coding will not, however, be sufficient for all purposes. It is available on an Internet site, and viewable using a web interface.

Introduction

Alors que de nouveaux outils de communication écrite synchrone, tels que les salons de discussion, la messagerie instantanée et le texto, connaissent un essor indéniable depuis quelques années, leurs spécificités linguistiques, pourtant reconnues, restent encore peu étudiées dans le détail. Il est vrai que le manque de ressources les concernant, tout au moins pour la langue française, ne fait rien pour en faciliter l'étude. Nous nous intéressons ici plus particulièrement à la langue du tchat, la « langue tchatée », produite à l'aide d'outils de tchat tels que les salons de discussion ou les différents logiciels de messagerie instantanée, laissant ainsi de côté la langue des textos, que nous supposons assez différente. Aujourd'hui, toute étude de la langue du tchat passe par la constitution d'un corpus, mais les contraintes de temps font que, bien souvent, ce dernier est assez réduit, aussi bien du point de vue de la longueur, que du nombre d'utilisateurs impliqués, ou encore des thèmes abordés; ce qui peut amener à s'interroger sur la portée réelle des résultats obtenus. Pourtant, ces nouveaux outils de

communication, moins normatifs que l'écrit traditionnel, offrent une opportunité de jeter un regard nouveau sur la langue écrite.

Au printemps 2004, au cours d'un stage de M2R[1] portant sur la traduction automatique de tchat (Falaise, 2004), un important corpus de français tchaté a été collecté, afin d'évaluer les difficultés posées par la la langue du tchat à son traitement automatique. Nous souhaitons contribuer à l'étude de cette langue en mettant ce corpus à disposition. Après un bref aperçu des principes du tchat et des caractéristiques du « français tchaté »[2], nous présenterons donc ce corpus, depuis sa méthode de collecte originale, jusqu'à son mode de diffusion.

1 Qu'est-ce que le tchat ?

1.1 Les outils de tchat

Nous considérons, à la suite de (Latzko-Toth 2001), que les dispositifs tels que les salons de discussion et les messageries instantanées peuvent être regroupées sous l'appellation « d'outils de tchat ». Les salons de discussion se confondent généralement avec le principal protocole sur lequel ils s'appuient, à savoir le protocole IRC. Il existe de nombreux réseaux IRC, et chaque réseau, se divise en canaux (ou salons) indépendants les uns des autres, et souvent associés à un thème précis. La messagerie instantanée, qui se distingue des salons de discussion, ouverts à tous, par son caractère privé, repose quant à elle sur un grand nombre de protocoles, chaque réseau ayant recours au sien.

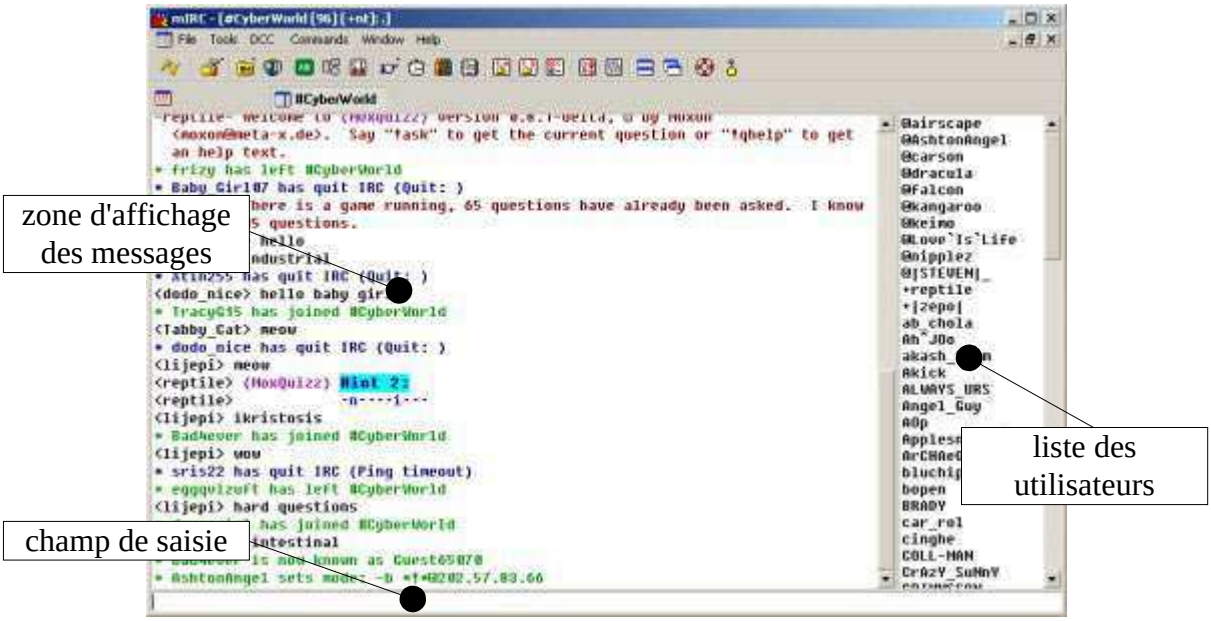

Figure 1 : Une session de tchat dans mIRC, le plus connu des clients de tchat pour le réseau IRC.

Ces dispositifs se distinguent notamment d'autres outils informatiques, tels que le forum ou le courriel, par leur caractère synchrone; en cela, ils se rapprochent plus du texto. De plus, comme ce dernier, le tchat est *volatil*. Le contenu d'une session de tchat, à l'instar d'une conversation verbale, n'a pas vocation à être enregistrée. Ainsi, lorsqu'un utilisateur se

[1] Seconde année de master recherche.

[2] Pour reprendre l'expression de (Pierozak, 2003).

connecte, il est dans l'ignorance totale de ce qui a été dit avant son arrivée, de même qu'il ne pourra pas savoir ce qui se dira après sa déconnexion. Mais à la différence du texto, les outils de tchat proposent aux utilisateurs des espaces communs, qui rendent aisée la communication entre de nombreuses personnes, là où le texto peut difficilement mettre en relation plus de deux personnes. De plus, il faut garder à l'esprit que les textos sont facturés au message, ce qui incite l'utilisateur à être synthétique et à en envoyer le moins possible, alors qu'au contraire le tchat, moins contraignant de ce point de vue, autorise toutes les digressions.

1.2 La langue du tchat

Les principales caractéristiques de la langue du tchat sont présentées, notamment, dans (Pierozak, 2003) et (Guimier de Neef & Véronis, 2004), et on n'en exposera donc ici que les grands principes.

Le « français tchaté », ou « clavardage » pour reprendre le terme québécois, se caractérise en particulier par une syntaxe proche de la langue parlée, et surtout par sa graphie originale. Loin d'être une limitation, le caractère écrit des conversations de tchat semble en effet en être l'un des principaux attraits (Herring, 1999, et Latzko-Toth, 2001). En fait, dans la langue du tchat, la graphie d'un lexème semble plus relever de la fantaisie de son auteur que de la norme orthographique, selon un processus de création lexicale permanente que (Pierozak, 2003) qualifie de « ludogénèse », et suffisamment souple pour permettre à certains utilisateurs de développer leur propre « voix » graphique.

Comme le souligne (Pierozak, 2003), la syntaxe des énoncés de tchat tient beaucoup de l'oral. Entre autres phénomènes propres à la langue parlée, les topicalisations y sont fréquentes, ainsi que les constructions du type *situation* + *thème* + (*rhème*). En outre, pour éviter les messages trop longs, les usagers découpent souvent leurs messages en propositions, ce qui n'est pas sans rappeler le découpage en groupes prosodiques (Falaise, 2004).

Pour caractériser les spécificités lexicales du tchat par rapport à l'écrit « standard », on pense bien entendu aux émoticons (« :-) », « ^^ », etc.) et aux abréviations, fréquentes en tchat, et parfois spécifiques à ce mode de communication (comme « lol » , « mdr », « tlm », etc.). On relève par ailleurs fréquemment une graphie que l'on pourrait qualifier de phonétique (« salut les zamis »), et qui sert souvent à transcrire des variantes phonolgiques (« kikoo » pour « coucou », « oki » pour « okay », etc.) : il s'agit d'une modification de la graphie d'un mot, destinée à lui donner une prononciation légèrement différente de la norme. Parfois, ces variantes phonologiques correspondent à des allongements vocaliques, comme dans « kikooooooo » par exemple. Comme l'a fait remarquer (Guimier de Neef & Véronis, 2004), on voit réapparaître en tchat des phénomènes de créativité phonético-graphique que l'on pensait réservés aux systèmes d'écriture anciens (hiéroglyphes égyptiens, anciens sinogrammes, glyphes mayas, entre autres...) : l'insertion de signes autonomes, porteurs d'une signification propre, dans des mots, d'après leur valeur phonétique et sans tenir compte de leur valeur sémantique. Par exemple « 2m1 », « 2main » et « dem1 »[3] pour « demain ». On peut aussi relever des graphies résultant d'une créativité sémantico-graphique, telles que « Micro$oft »[4] pour « Microsoft », dans lesquelles on insère un signe possédant des valeurs lexicale et sémantique propres, qui sont cette fois toutes deux conservées dans la graphie ainsi formée. Ces phénomènes de créativité graphique, bien que bien connus et pour certains aussi anciens que l'écriture elle-même, ne sont pas présents dans la langue écrite normée moderne, et sont donc généralement négligés en TALN.

[3] Respectivement 163, 31 et 137 occurrences de ces formes dans notre corpus.

[4] 14 occurrences dans notre corpus.

La figure 2 donne une idée générale de l'importance de ces divergences lexicales, à défaut d'une analyse plus fine. On constate que les deux tiers environ des « mots » (caractères entre deux espaces, y compris émoticons) sont correctement orthographiés, et que les mots involontairement mal orthographiés (« orthographe mal formé ») sont rares. Par contre, un nombre significatif de mots relèvent de la graphie phonétique (« orthographe phonétique ») décrite au paragraphe précédent. Le recours aux abréviations, aux émoticons, aux onomatopées, ainsi que les références aux autres utilisateurs, se révèlent aussi assez courants. Enfin, on relève quelques cas de xénismes (des anglicismes en l'occurrence) et de fusions de mots (« jme demande », « jte dis », « ça mva », etc.).

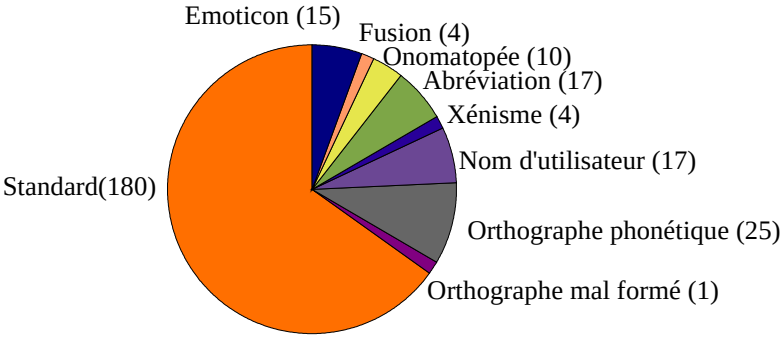

Figure 2 : répartition des principaux phénomènes lexicaux dans un ensemble de 77 messages déterminés aléatoirement au sein du canal #18-25ans; nombre d'occurrences entre parenthèses.

Il faut souligner néanmoins que malgré toutes ces divergences par rapport à l'écrit « standard », la finalité du tchat demeure le dialogue, et les tchateurs sont soucieux, jusqu'à un certain point, de la clarté de leur messages. En témoignent les corrections effectuées *a posteriori* par les tchateurs eux-même, généralement lorsqu'un mot mal orthographié peut être confondu avec l'un de ses homophones hétérographes (« s/pere/paire/ », vu sur le canal #c++).

Ainsi, l'un des intérêts scientifiques de l'étude de la langue du tchat tient au fait qu'elle vient remettre en cause certains *a priori* fréquents en TALN, comme par exemple l'approche scholastique de la notion de grammaticalité ou encore le caractère fermé du lexique.

2 Collecte du corpus

2.1 Ethique

La collecte d'un corpus tchaté ne va pas sans soulever certaines questions éthiques. En effet, même en nous limitant aux salons de discussion, publics, il s'agit tout de même d'enregistrer des conversations en principe éphémères. Et à moins de créer un canal dédié à la collecte d'un corpus, il n'est pas possible de prévenir tous les utilisateurs du fait qu'ils vont être enregistrés.

Nous avons donc choisi de constituer ce corpus à partir d'enregistrements de salons de discussion, librement consultables sur Internet, plutôt que d'enregistrer nous-même des sessions de tchat. Un autre avantage de cette méthode est qu'elle permet de récupérer en une seule fois une grande quantité de données.

2.2 Droit d'auteur

Cette approche ne lève pas toutefois tous les doutes qui se posent au niveau des droits d'auteur. Pour certains canaux, représentant environ 15% des messages du corpus, les conditions d'utilisation[5] stipulent que les enregistrements sont consultables par tous et reproductibles : « tout utilisateur accepte que les propos qu'il tient sur les canaux officiels puissent être visibles et transmis ». Rien n'est précisé pour les autres canaux. Le code de la propriété intellectuelle (CPI) se montre peu clair en ce qui concerne les dialogues publics anonymes. En admettant que les dialogues de tchat soient protégés par le droit d'auteur, ce qui n'est pas évident au vu de l'article L.112-2 du CPI[6], il semble difficile de considérer chaque message comme une œuvre à part entière, à moins de considérer des textes tels que « :-) », « salut » ou « ouaip » comme des œuvres originales. De plus, dans un tel contexte de dialogue, les messages peuvent difficilement être compris sans se référer à leur contexte. Par conséquent, on pourrait plutôt considérer chaque canal de tchat comme une œuvre collective[7] anonyme[8]. Selon l'article L.113-6 du CPI, les œuvres anonymes sont gérées par leur éditeur, c'est à dire en l'occurrence, soit le responsable du serveur de tchat, soit celui du site sur lesquels les enregistrement de tchat sont publiés. Ce problème, et surtout sa solution, nous étant apparus assez tard, nous cherchons actuellement à déterminer laquelle de ces deux personnes est, aux yeux de la loi, dépositaire des droits d'auteur, afin de la contacter pour obtenir une autorisation de publication en bonne et due forme.

2.3 Collecte

Notre corpus de langue tchatée est constitué à partir des enregistrements disponibles sur le site http://www.botstats.com. Ce site publie les résultats du service web Botstats, dédié aux canaux de tchat, et qui permet notamment la tenue de statistiques et l'archivage des discussions pour une durée maximale de trois mois. Ce service est utilisé par quelques centaines de canaux du réseau IRC EpikNet. La publication des enregistrements est un choix de la part du créateur du canal; et ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux (une centaine) ont décidé de les rendre accessibles à tous, et ce sont les enregistrements de ces canaux que nous avons regroupés au sein du corpus.

Les enregistements, consultables sous forme de pages HTML sur Internet, sont tout d'abord extraits à l'aide d'un « aspirateur de sites », puis les fichiers HTML obtenus sont convertis automatiquement en XML grâce à des expression régulières, afin de pouvoir être exploités.

2.4 Format des données

L'activité d'un canal de tchat peut être représentée par une succession de messages, produits par différents auteurs. Outre les messages « normaux », rédigés par un auteur humain à destination de lecteurs humains, il faut distinguer quelques cas particuliers :

---

5 http://www.epiknet.org/legal/

6 Cet article, qui décrit ce qui est protégé par le droit d'auteur, n'est toutefois pas restrictif.

7 Au sens de l'article L.113-2 du CPI.

8 Au sens de l'article L.113-6 du CPI.

- les commandes, qui sont destinées au serveur (afficher la liste des utilisateurs par exemple) ou à un robot (sur un canal de tchat, un robot, ou « bot », a le statut d'utilisateur et peut par exemple intervenir pour rappeler le thème du canal, donner l'heure, gérer des jeux, mener des dialogues de type Eliza[9], etc.), et qui appartiennent à un langage formel;

- les messages pré-enregistrés, déclenchés à l'aide de raccourcis clavier ou lors de certains événements (déconnexion de l'utilisateur par exemple), qui ne relèvent pas du même contexte pragmatique;

- les messages envoyés par des robots;

- les événements, notifiés par le serveur (quelqu'un vient se connecter, de changer de surnom, etc...).

Le corpus est codé en XML. L'élément racine <*log*>, peut avoir quatre types d'éléments-fils :

- l'élément <*commentaire*>, dont le contenu est un commentaire sur le canal;

- l'élément <*message*>, comportant un message envoyé par un utilisateur, humain ou non, et destiné à être lu par les autres utilisateurs humains;

- l'élément <*commande*>, comportant une commande destinée au serveur;

- l'élément <*evenement*>, dont le contenu est un événement notifié par le serveur.

Les éléments <*message*>, <*commande*> et <*evenement*> possèdent des attributs *date* et *heure*. Les éléments <*message*> et <*commande*> comportent en outre un sous-élément <*auteur*>, contenant le surnom de l'utilisateur ayant produit le message, ainsi que des attributs indiquant son type, humain ou robot. <*evenement*> comporte quant à lui des sous-éléments précisant le type d'événement et simplifiant leur traitement automatique.

Le corpus, encodé automatiquement, respecte ces spécifications, à deux exceptions près. La valeur des attributs *type* (utilisateur humain ou robot), qui ne peut être déterminée automatiquement, puisque rien ne distingue formellement les messages d'un humain de ceux d'un robot, n'est pour l'instant pas renseignée. De plus, un certain nombre de commandes n'ont pas été reconnues par les expressions régulières chargées de les identifier, et ces dernières devront être affinées. En effet, les commandes peuvent généralement être reconnues par l'expression régulière ^!.+ , correspondant à une ligne débutant par un point d'exclamation, mais certains canaux proposent des commandes supplémentaires, dont la syntaxe est différente, et qui ne seront par conséquent pas détectées comme telles. Inversement, il arrive parfois, bien qu'assez rarement, qu'un message normal débute par un point d'exclamation : ce message sera alors considéré à tort comme une commande. Il convient donc d'élaborer une nouvelle expression de recherche pour chaque canal, en tenant compte de la liste des commandes et de leur syntaxe exacte.

[9] « ELIZA est un célèbre programme informatique écrit par Joseph Weizenbaum, qui simulait un psychothérapeute rogérien en reformulant la plupart des affirmations du "patient" en questions, et en les lui posant. » (Wikipédia, http://fr.wikipedia.org/wiki/ELIZA)

	<log xml:lang="fr">
	<commentaire>Exemple</commentaire>
A> soir tlm	<message date="29/03/2004" heure="15:13">
	<auteur type="humain">A</auteur>
	soir tlm
	</message>
A> kikooooooooB :*)	<message date="29/03/2004" heure="15:20">
	<auteur type="humain">A</auteur>
	kikooooooooB :*)
	</message>
B> kikoooA :)	<message date="29/03/2004" heure="15:20">
	<auteur type="humain">B</auteur>
	kikoooA :)
	</message>
<C vient de se connecter>	<evenement date="29/03/2004" heure="15:25">
	<connexion>
	<utilisateur>C</utilisateur>
	</connexion>
	C vient de se connecter
	</evenement>
	</log>

Exemple 1 : exemple de discussion et code XML correspondant.

3 Première évaluation des résultats

3.1 Quantification du corpus

D'un point de vue quantitatif, la somme de données collectées est assez considérable : 4 192 033 messages, couvrant environ 3 mois de conversations sur 105 canaux de tchat. Si l'on considère un mot comme une suite de caractères délimitée par les signes de ponctuations traditionnels (cette définition n'est pas forcément la plus adaptée à la langue du tchat, mais est acceptable en première approche), alors le corpus comporte 23 011 876 mots, soit une moyenne d'environ 5,5 mots par message. De ce point de vue, ce corpus apparaît sans commune mesure avec l'existant, et ce d'autant plus qu'on peut l'étendre en continu, au fur et à mesure que de nouveaux enregistrements sont générés par le service Botstats. A titre de comparaison, le plus important corpus auquel nous ayons eu accès est le corpus d'italien tchaté constitué par la société Eulogos (Eulogos, 2001), qui comporte 849 510 mots.

3.2 Evaluation qualitative

Les thèmes abordés par les canaux sont variés, et vont du tchat généraliste où l'on discute de tout et de rien, au tchat spécialisé dans les problèmes de programmation, ou encore les débats concernant l'actualité. On relève aussi des différences d'ordre pragmatique. En plus des traditionnels bavardages, certains canaux sont plus ou moins dédiés aux jeux (pendu, quizzes), alors que dans d'autres la conversation est alimentée par des dépêches AFP. D'autres enfin sont consacrés à des discutions techniques, sous forme de question/réponse, par exemple sur un canal consacré aux questions de programmation.

Certains canaux semblent à première vue assez originaux sur le plan linguistique. Ainsi on peut constater en comparant les figures 3 et 4 que le canal #edelweiss comporte peu de formes pour sa taille (42% de moins que #ffparadise, de taille pourtant plus réduite). D'un point de vue plus général, il semble qu'un corpus de tchat comporte nettement plus de formes qu'un

corpus écrit « standard » ou oral équivalent, quand on compare notre corpus avec ceux décrits dans (Gendner V. & Adda-Decker M., 2002).

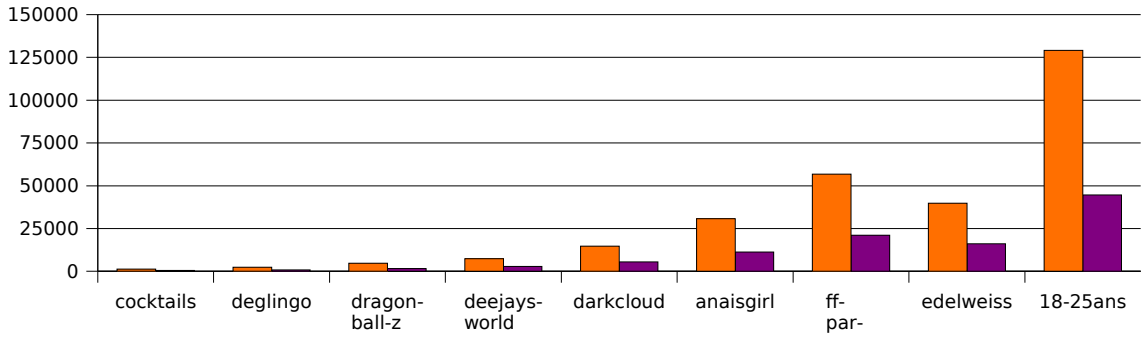

Figure 3 : nombre de formes (barre de gauche), et nombre de formes présentes au moins deux fois (barre de droite), dans quelques canaux du corpus.

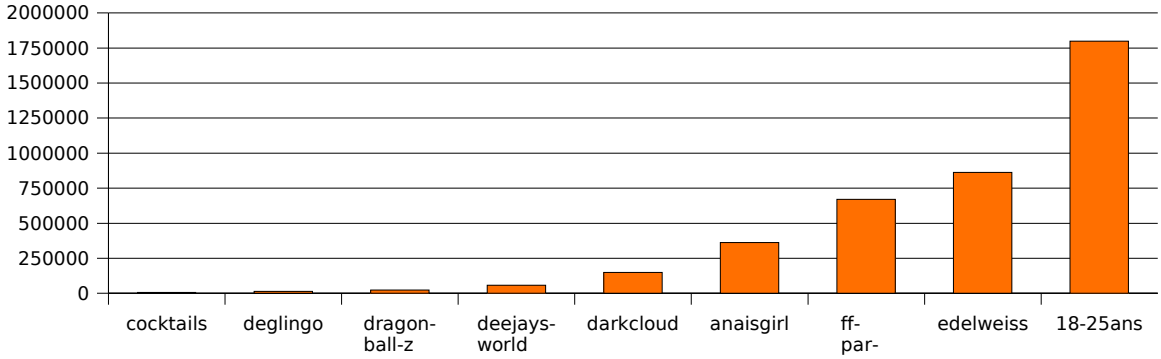

Figure 4 : nombre de mots dans quelques canaux du corpus.

Toujours par comparaison avec les corpus écrit et oral de (Gendner V. & Adda-Decker M., 2002), on peut constater que le nombre de formes présentes plusieurs fois, par rapport au nombre total de formes, est beaucoup plus faible en tchat (35% des formes pour le canal #18-25ans) que pour l'écrit « standard » (62%) et l'oral (70%).

3.3 L'avenir

Dans un premier temps, il est nécessaire d'effectuer un classement des canaux, en fonction de leur thème mais aussi de leurs spécificités pragmatiques (type d'interaction et d'interacteur[10]), afin de pouvoir ensuite sélectionner ceux qui correspondent le mieux à ce que l'on veut étudier. On ne s'attend pas, en effet, à ce qu'un canal de jeu ait les mêmes propriétés qu'un canal de conversation classique. L'encodage doit aussi être amélioré, comme décrit en 2.4, de façon à correspondre aux spécifications; il s'agit d'un travail semi-manuel relativement important, en particulier en ce qui concerne la notation du caractère humain ou non de l'auteur du message (attribut *type*).

Enfin, une annotation lexicale plus fine est envisageable, permettant d'identifier, et par conséquent de quantifier, les différentes particularités graphiques de la langue du tchat, comme les émoticons, déformations graphiques, corrections orthographiques, etc. fréquemment relevés dans l'étude de la langue tchatée, de façon plus fine que ce qui est présenté en première partie. Toutefois ce dernier traitement, manuel, est très lourd à mettre en œuvre, et ne peut pas concerner tout le corpus. Il est aussi possible d'obtenir des résultats

[10] Humain ou robot.

intéressants pour caractériser un canal ou un utilisateur, à partir de quelques centaines de mots prélevés aléatoirement, comme nous en avons donné un bref aperçu en 1.2.

4 Mise à disposition du corpus

Notre corpus est consultable en ligne[11], grâce à une interface web. Afin de permettre sa consultation dans des conditions raisonnables, celui-ci a été transféré dans une base de données MySQL; un script PHP regénère à la volée le code XML correspondant à la partie du corpus sélectionnée dans l'interface. Ce code XML est associé à une feuille de style XSL permettant une visualisation simple des données au sein de l'interface, à condition d'utiliser un navigateur supportant ce format[12]. L'intérêt de l'utilisation dynamique d'une feuille XSL est que le code XML reste disponible, par exemple lorsque l'on demande au navigateur d'afficher le code source de la page.

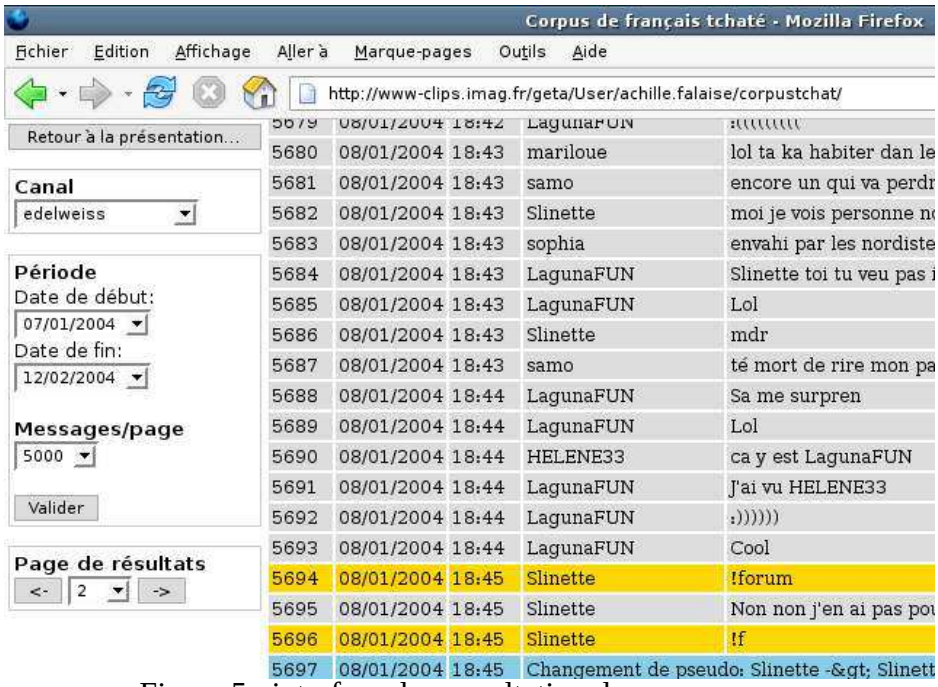

Figure 5 : interface de consultation du corpus.

A terme, un système d'enregistrement et d'authentification des utilisateurs de ce système sera mis en place, afin de restreindre son utilisation au seul monde scientifique.

Conclusion

Ce corpus, de par son étendue, tant du point de vue du nombre de mots (plus de 23 millions), que du nombre de canaux (105), est assez représentatif de la langue tchatée. Il peut ainsi, malgré certaines insuffisances, être utilisé en complément de corpus plus précis mais aussi plus restreints, dans le cadre de l'étude de la langue tchatée, ou encore pour évaluer

---

[11] http://www-clips.imag.fr/geta/User/achille.falaise/corpustchat/

[12] Ce format est supporté par Firefox 1.0, ainsi que par Internet Explorer 4.5+ sous Windows, après l'installation de la librairie MSXML pour les versions de Windows antérieures à Windows XP.

rapidement un outil de TALN dans ce cadre linguistique, ce pourquoi il était conçu à l'origine. C'est pourquoi nous pensons utile de le mettre à disposition de la communauté du TALN.

Progressivement, les nouveaux outils de communication écrite nous amènent à élargir notre conception de la langue écrite, et à reconsidérer certains principes du TALN qui semblaient acquis (grammaticalité des énoncés, lexique sous forme de listes, etc.). Nous pouvons constater, avec l'exemple de la langue tchatée, à quel point ces conceptions sont liées au caractère contraint de « l'écrit standard », et demandent à être élargies pour vraiment rendre compte des réalités cognitives à l'œuvre dans la langue.

Références

Eulogos (2001), « Corpus di conversazioni da chat-line in lingua italiana, da registrazioni effettuate nel primo trimestre 1998 »
http://www.intratext.com/X/ITA0192.HTM

Falaise (2004) : Premier pas vers une TA interactive pour le tchat, rapport de stage de master, Université Joseph Fourier, Grenoble, 63 pages.

Gedner V. & Adda-Decker M. (2002), Analyse comparative de corpus oraux et écrits français: mots, lemmes et classes morpho-syntaxiques, *Actes des XIVe Journées d'Etude sur la Parole*, Nancy.

Guimier de Neef E. & Véronis J. (2004) : « 1 pw1 sr la kestion ;-). », Journée d'étude de l'ATALA, *Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)*, Paris.

http://www.up.univ-mrs.fr/~veronis/je-nfce/resumes.html#1pw1

Herring S. (1999), « Interactional coherence in CMC », *Journal of computer-Mediated Communication*, Vol. 4, n°4.

Latzko-Toth G. (2001), « Un dispositif construit par ses utilisateurs ? Le rôle structurant des pratiques de communication dans l'évolution technique de l'Internet Relay Chat », *Actes du IIIème colloque international sur les usages et services des télécommunications*, pp. 556-564.
http://grm.uqam.ca/textes/Latzko_ICUST2001.pdf

Pierozak I. (2003), Le "français tchaté" : un objet à géométrie variable ?, *Langage et Société*, n° 104, pp. 123-144.

Pujade L. (2001), L'écrit sur internet, mémoire de maîtrise, Toulouse, 179 pages.

Shortis, T. (2000), *The Language of ICT*, London, Routledge.